Variabili aleatorie discrete

Giovanni M. Marchetti

Statistica Capitolo 5 — Corso di Laurea in Economia 2015-16

Variabili aleatorie

Una variabile aleatoria è simile a una variabile statistica

- Una variabile statistica è definita da un insieme di modalità cui sono associate delle frequenze
- Una variabile aleatoria (o casuale) è definita da un insieme di modalità cui sono associate delle probabilità

Distinzione

Le variabili aleatorie possono essere

1. discrete
2. continue

Variabili discrete

Le modalità sono numeri interi.

– Lancio due monete: il numero di teste X è una v.a. discreta

x	0	1	2	Totale
$p(x)$	$1/4$	$1/2$	$1/4$	1

Notazione: $p(x) = P(X = x) = P(\text{numero di teste} = x)$

Spiegazione

Perché $p(1) = 1/2$? Guardate la tabella dei risultati

$$\begin{array}{cc|cc} & & T & C \\ \hline T & & TT & TC \\ C & & CT & CC \end{array}$$

Quindi

$$p(1) = P(TC \cup CT) = P(T \cap C) + P(C \cap T) = \tfrac{1}{2}\tfrac{1}{2} + \tfrac{1}{2}\tfrac{1}{2} = \tfrac{1}{2}$$

In generale

Una variabile aleatoria discreta ha una funzione di massa di probabilità $p(x)$ tale che

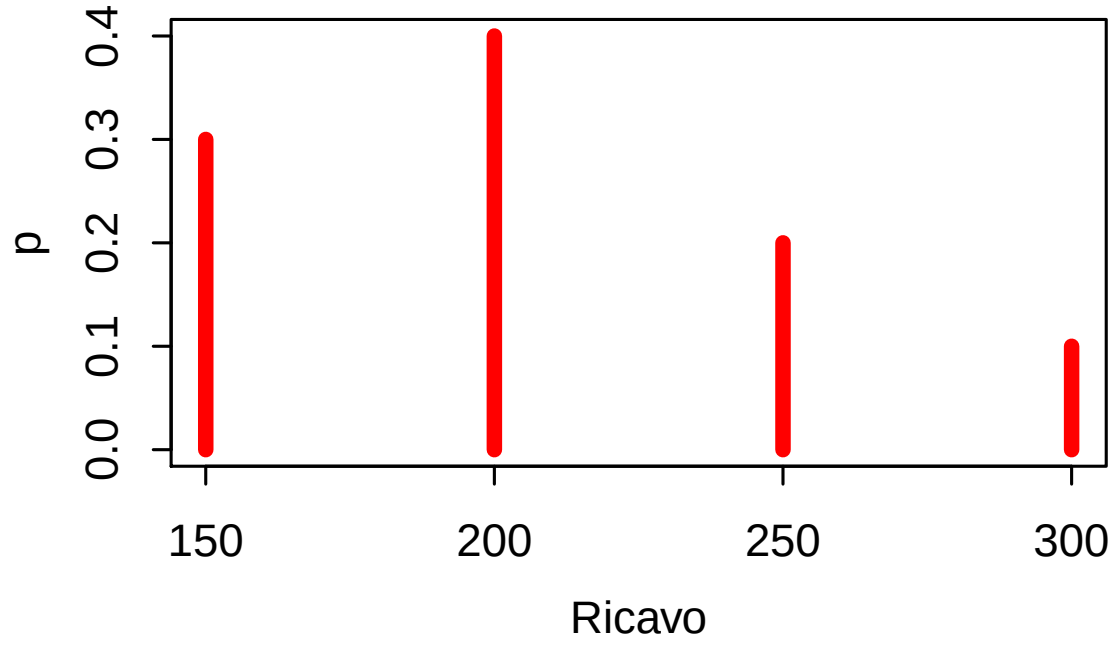
- $p(x) \geq 0$ (è la probabilità che X sia uguale a x)
- $\sum_x p(x) = 1$ (è la legge dell'inevitabilità)

La funzione di massa di probabilità definisce la distribuzione di probabilità di X.

Esempio

Il ricavo previsto X (in migliaia di euro) della vendita di un immobile sia descritto dalla distribuzione

x	150	200	250	300	Totale
$p(x)$	0.3	0.4	0.2	0.1	1.0



Funzione di ripartizione

Se si calcolano le probabilità cumulate si ottiene la funzione di ripartizione.

x	150	200	250	300
$p(x)$	0.3	0.4	0.2	0.1
$F(x)$	0.3	0.7	0.9	1.0

Esprime la probabilità che X non superi un dato valore

Definizione:

$$F(x) = P(X \leq x)$$

Esempio: $F(200) = P(\text{ricavo sia al massimo } 200) = 0.7$

Media di una variabile aleatoria

Supponiamo che X abbia k modalità $x_1, \ldots, x_k$ e denotiamo le probabilità associate con

$$p_i = p(x_i)$$

La media di una variabile aleatoria discreta si calcola con

$$\mu = \sum_{i=1}^{k} x_i \, p_i$$

e si chiama valore atteso. È del tutto analogo alla media di una variabile statistica.

Calcolo

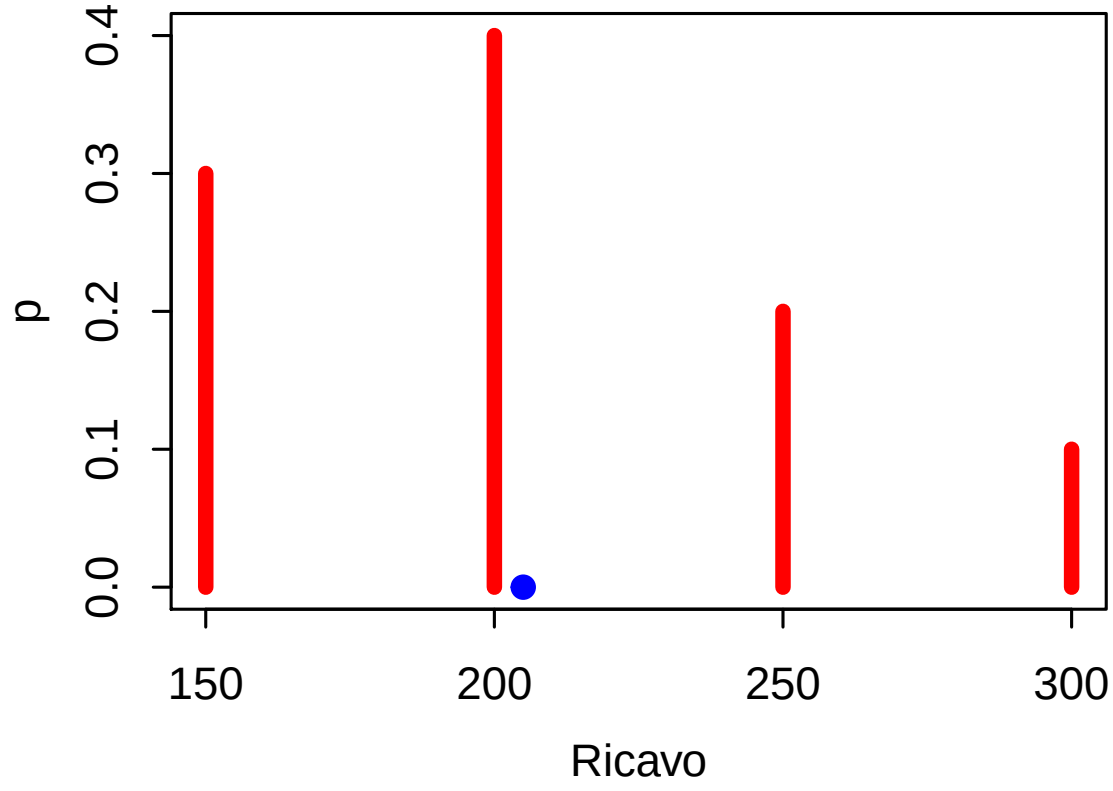
x_i	150	200	250	300	Totale
p_i	0.3	0.4	0.2	0.1	1.0
$x_i\ p_i$	45	80	50	30	205

Il ricavo atteso dalla vendita è

$$\mu = (150)(0.3) + (200)(0.4) + (250)(0.2) + (300)(0.1) = 205$$

Il valore atteso μ si indica anche con l'operatore $E(X)$.

Distribuzione e media



Varianza

La varianza di una variabile aleatoria discreta X si calcola con la stessa regola delle variabili statistiche

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \, p_i$$

Spesso è utile calcolarla con la formula alternativa

$$\sigma^2 = \sum_{i=1}^{k} x_i^2 \, p_i - \mu^2$$

La varianza di X si indica anche con l'operatore $\mathrm{var}(X) = \sigma^2$

Calcolo

x_i	150	200	250	300	Totale
x_i^2	22500	40000	62500	90000	
p_i	0.3	0.4	0.2	0.1	1.0
$x_i^2\ p_i$	6750	16000	12500	9000	44250

Quindi la varianza del ricavo è

$$\sigma^2 = \sum_{i=1}^{k} x_i^2\ p_i - \mu^2 = 44250 - 205^2 = 2225.$$

Deviazione standard

La deviazione standard è la radice quadrata della varianza: $\sigma = \sqrt{\sigma^2}$.

σ misura quanto è affidabile la media μ.

La deviazione standard del ricavo previsto è

$$\sigma = \sqrt{2225} \simeq 47$$

Mi aspetto un ricavo medio di 205000 euro con una variabilità di 47000 euro.

Coefficiente di variazione

Il coefficiente di variazione è $CV = \sigma / \mu$

Serve per avere una misura di variabilità relativa (cioè senza unità di misura) confrontabile su distribuzioni diverse.

Esempio

Il ricavo previsto (in migliaia di euro) Y dalla vendita di un'auto è descritto dalla distribuzione seguente

y_i	5	6	7	8	9	Totale
p_i	0.2	0.2	0.2	0.2	0.2	1

Come confrontare la variabilità dei due ricavi X e Y?

$\sigma_X = 47$ ($\mu_X = 205$), $\sigma_Y = 1.41$ ($\mu_Y = 7$)

$$CV_X = 47/205 = 0.23, \qquad CV_Y = 1.41/7 = 0.201$$

Grandezze funzioni di variabili aleatorie

X sia il numero previsto di giorni necessari per finire un progetto, con distribuzione

x_i	10	11	12	13	14	Totale
p_i	0.1	0.3	0.3	0.2	0.1	1

Per esercizio verificate che $\mu = 11.9$ giorni e $\sigma = 1.29$ giorni.

Se ci sono costi fissi per 25000 euro e un costo di 900 euro per ogni giorno di lavoro quant'è il costo totale C?

Per definizione è $C = 25000 + 900X$ e quindi è una variabile aleatoria.

Distribuzione del costo totale

$$C = 25000 + 900X$$

c_i	34000	34900	35800	36700	37600	Totale
p_i	0.1	0.3	0.3	0.2	0.1	1

Per calcolare media e deviazione standard non occorre rifare il calcolo: basta usare le formule

$$\begin{aligned}
\mu_C &= E(25000 + 900X) = 25000 + 900E(X) \\
&= 25000 + (900)(11.9) = 35710 \text{ euro} \\
\sigma_C^2 &= \text{var}(25000 + 900X) = 900^2 \text{var}(X) = 1044900 \\
\sigma_C &= \sqrt{1044900} \simeq 1022 \text{ euro}
\end{aligned}$$

Una distribuzione discreta fondamentale

Si dice che X è una variabile di Bernoulli se assume solo due valori

- $x = 1$ detto successo
- $x = 0$ detto insuccesso

con probabilità rispettivamente p e $1 - p = q$. Cioè

x	0	1	Totale
$p(x)$	q	p	1

L'esperimento associato si dice prova di Bernoulli.

Esempi

- X = risultato di un processo che produce schede telefoniche ($x = 1$ significa che il pezzo è difettoso)

x	0	1	Totale
$p(x)$	0.999	0.001	1

- X = risultato del lancio di una moneta ($x = 1$ significa che esce testa)
- X = restituzione di un mutuo ($x = 1$ significa che il mutuo è restituito)

Prove di Bernoulli

- Se si fanno n prove di Bernoulli indipendenti si hanno n variabili aleatorie di Bernoulli

$$X_1, X_2, \ldots, X_n$$

ciascuna con la stessa distribuzione (stessa probabilità di successo p) e indipendenti (cfr. dopo)

- Tipicamente alla fine si studia la distribuzione del numero di successi nelle n prove

$$S = X_1 + \cdots + X_n$$

Esempio: test a crocette

- Un test contiene 2 domande (poi complichiamo) in cui entrambe hanno 5 possibili risposte A, B, C, D di cui una sola è giusta
- Un robot estrae a sorte le risposte alle due domande.
- Qual è la probabilità che risponda a $s = 0, 1, 2$ domande?

Struttura del problema

- Si tratta di due prove di Bernoulli (in ognuna il robot può avere un successo o un insuccesso)
- Se si sceglie a caso la probabilità di successo è $p = 1/4 = 0.25$
- Le due prove sono indipendenti (il robot non ha conoscenze e non impara)
- Il numero di successi è $S = X_1 + X_2$

Soluzione

La distribuzione del numero di successi S in due prove è

s	0	1	2	Totale
$p(s)$	q^2	$2pq$	p^2	1

e prende il nome di Binomiale.

Perché? Perché le probabilità derivano dallo sviluppo del binomio $(q+p)^2$

- Poiché $p = 1/4$, le probabilità di $s = 0, 1, 2$ successi del robot sono

s	0	1	2	Totale
$p(s)$	$9/16$	$6/16$	$1/16$	1

- Se si passa il test con un punteggio di almeno 1, il robot ha una probabilità $7/16 = 43.7\%$ di passare il test.

Successi in 3 prove di Bernoulli

- Il robot prova un test con 3 domande a crocette. Stessa strategia.
- Quali sono le probabilità di $s = 0, 1, 2, 3$ successi?
- Soluzione: la distribuzione del numero di successi S in 3 prove è

s	0	1	2	3	Totale
$p(s)$	q^3	$3q^2p$	$3qp^2$	p^3	1

- La distribuzione di chiama ancora Binomiale perché le probabilità derivano da $(q + p)^3$.

- Il robot questa volta ha meno chances di passare il test.
- Sostituendo

s	0	1	2	3	Totale
$p(s)$	$27/64$	$27/64$	$9/64$	$1/64$	1

- La probabilità di passare il test (con almeno 2 punti su 3) è solo $10/64 \simeq 0.15$

Successi in n prove di Bernoulli

In generale si può calcolare la probabilità di s successi in n prove di Bernoulli indipendenti

Formula della Binomiale

Per $s = 0, 1, 2, \ldots, n$

$$P(S = s) = \binom{n}{s} p^s q^{n-s} = (1) \quad (2) \quad (3)$$

(1) Quante sono le n-uple con s successi

(2) (prob successo)$^{\#successi}$

(3) (prob insuccesso)$^{\#insuccessi}$

Esempi

- Probabilità che in 10 lanci di una moneta equa si ottengano 2 teste. Si ha $n = 10$, $p = 1/2$

$$p(2) = \binom{10}{2}(0.5)^2(0.5)^8$$
$$= \frac{10 \cdot 9}{2}(0.5)^{10} = 45/1024 = 0.0439$$

- Probabilità che in 10 lanci di una moneta si ottengano almeno 2 teste?

  $p(2) + p(3) + \cdots + p(10)$ giusto, ma non è conveniente!

  $1 - p(0) - p(1)$ giusto e conveniente

Problema

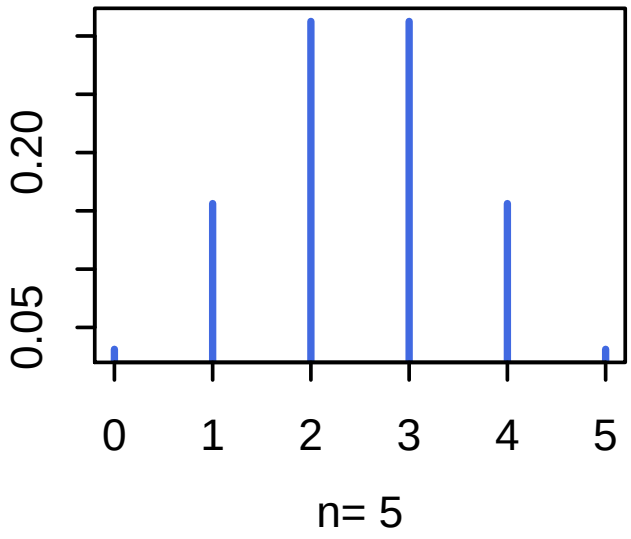
- Una ditta accetta un lotto di pezzi se un campione casuale di 20 pezzi non contiene più di un difettoso.

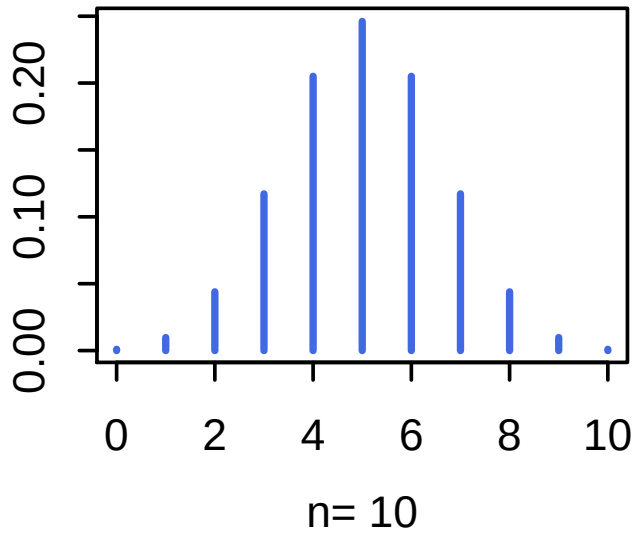
  Se la probabilità di difettoso è $p = 0.1$ qual è la probabilità che la ditta accetti il lotto?
- Risposta:

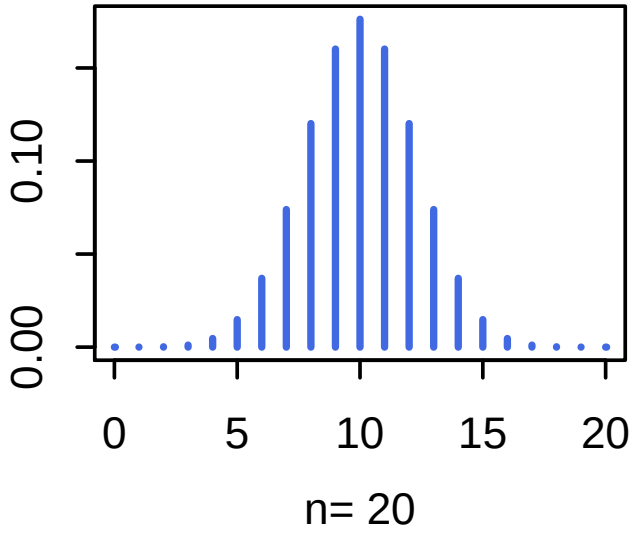
$$P(\text{accetta}) = p(0) + p(1) = \binom{20}{0} q^{20} + \binom{20}{1} q^{19} p$$
$$= (0.9)^{20} + 20(0.9)^{19}(0.1) = 0.39$$

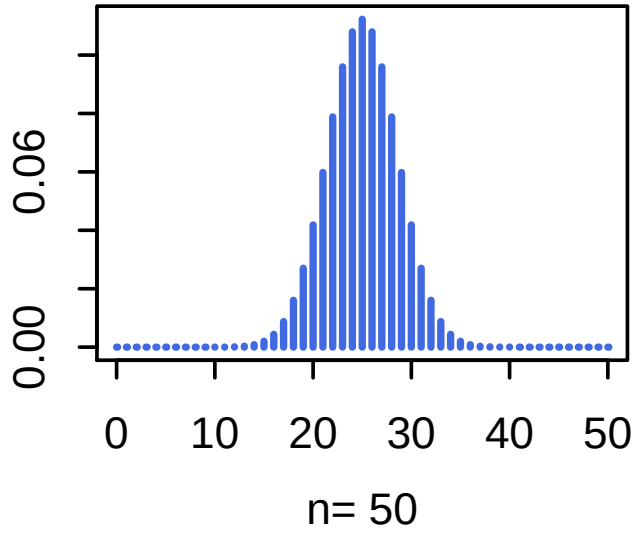
Figure

Se $p = 0.5$ la distribuzione Binomiale è simmetrica

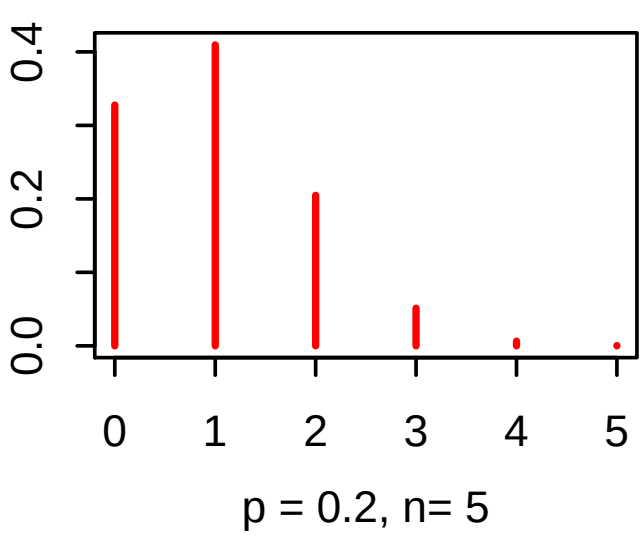


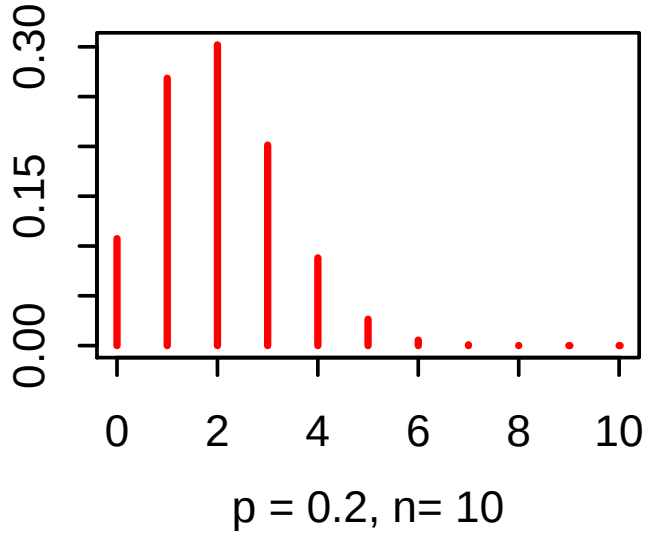


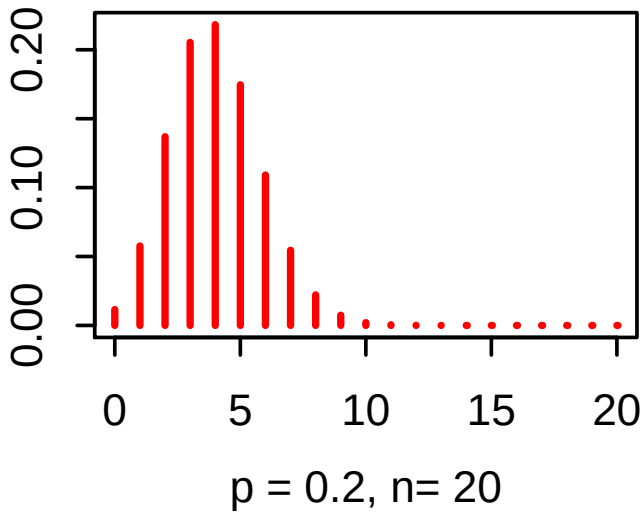


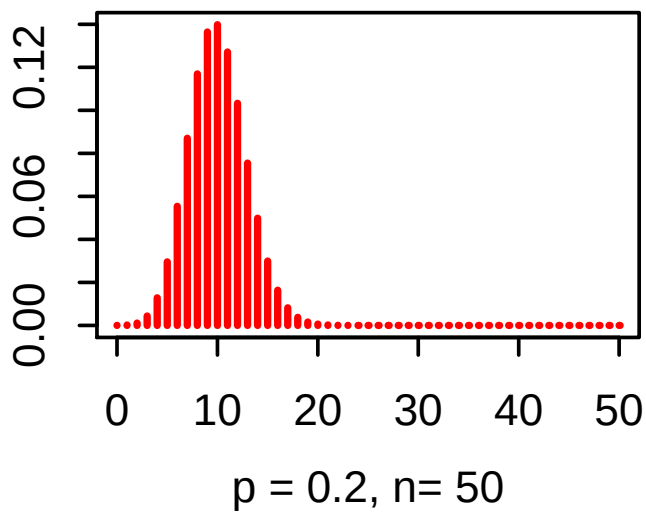


Se $p < 0.5$ la distribuzione Binomiale è asimmetrica

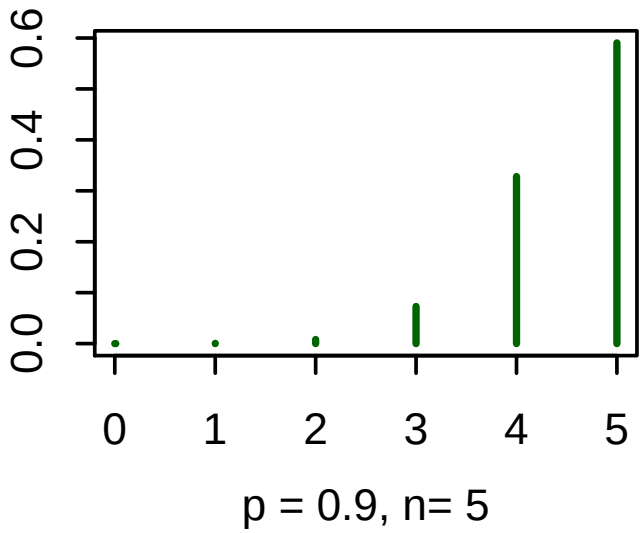


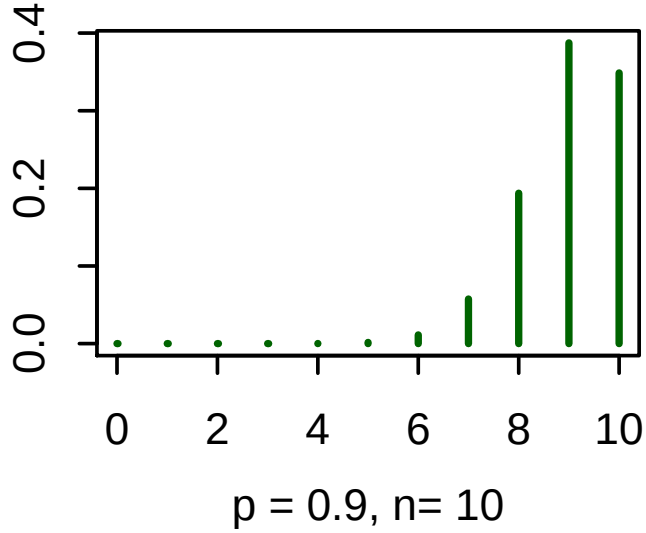


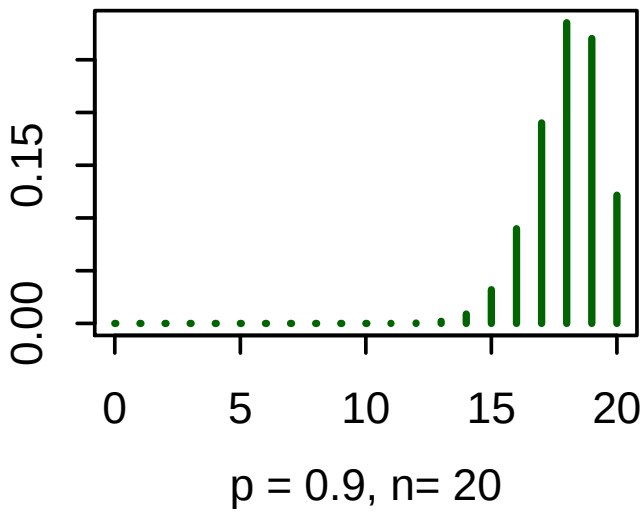


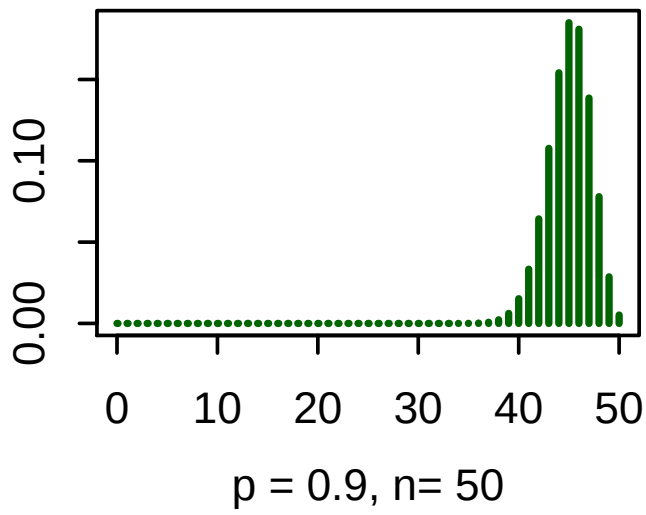


Se $p > 0.5$ la distribuzione Binomiale è asimmetrica









Proprietà della Binomiale

Se si fa una sola prova, cioè $n = 1$ la Binomiale coincide con la Bernoulli.

Valore atteso e varianza

Se S è Binomiale con probabilità di successo p e n prove

- $E(S) = np$
- $\text{var}(S) = npq$

Conseguenza:

Il valor medio della Bernoulli è $\mu = p$ e la varianza è $\sigma^2 = pq$.

Investimenti, rendimenti attesi, rischio

- Vogliamo investire una somma di 1000 euro in un titolo
- Il rendimento X è aleatorio e ha una distribuzione di probabilità $p(x)$
- La media e la deviazione standard sono fondamentali per valutare l'investimento
- $\mu_X 1000$ è il ricavo atteso e $\sigma_X 1000$ è il rischio

Confronto di due investimenti

x	$p(x)$
-0.05	0.4
$+0.20$	0.6
Totale	1
μ_X	0.1
σ_X	0.122

y	$p(y)$
$+0.00$	0.6
$+0.25$	0.4
Totale	1
μ_Y	0.1
σ_Y	0.122

Risulta: $\mu_X = (-0.05)(0.4) + (0.20)(0.6) = 0.1$.

Entrambi gli investimenti hanno

- lo stesso ricavo atteso (100 euro)
- lo stesso rischio (122 euro)

È bene diversificare?

Si investe

- una frazione α della somma di 1000 euro in X
- e una frazione $1 - \alpha$ in Y
- Il ricavo è $T = 1000[\alpha X + (1 - \alpha)Y]$
- Per rispondere alla domanda bisogna saper calcolare $E(T)$ e $\mathrm{var}(T)$.
- Questo richiede la conoscenza della distribuzione congiunta di X e Y

Distribuzione congiunta

- I due rendimenti X e Y sono collegati fra loro
- Le due variabili hanno una distribuzione doppia

$p(x,y)$	$y=0$	0.25	*Totale*
$x=-0.05$	0.1	0.3	0.4
$+0.20$	0.5	0.1	0.6
Totale	0.6	0.4	1.0

- $p(x,y) = P(X=x, Y=y)$ è la probabilità che i rendimento X sia x e Y sia y.
- Nei margini della tabella ci sono le distribuzioni separate di X e di Y

Indipendenza

- X e Y sono indipendenti se

$$p(x, y) = p(x)p(y)$$

- In questo caso non è così

Effettiva			
$p(x, y)$	$y = 0$	0.25	*Totale*
$x = -0.05$	0.1	0.3	0.4
$+0.20$	0.5	0.1	0.6
Totale	0.6	0.4	1.0

Indipendenza			
$\widehat{p}(x, y)$	$y = 0$	0.25	*Totale*
$x = -0.05$	0.24	0.16	0.4
$+0.20$	0.36	0.24	0.6
Totale	0.6	0.4	1.0

Covarianza

- Come per le variabili statistiche si misura la dipendenza lineare con la covarianza
- Per definizione

$$\sigma_{XY} = \operatorname{cov}(X, Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_y) p(x, y)$$

- Esiste una formula di calcolo alternativa

$$\sigma_{XY} = \sum_x \sum_y x\, y \cdot p(x, y) - \mu_X \mu_y$$

Esempio di calcolo

- Calcolo della media dei prodotti

$x \cdot y$	$y=0$	0.25
$x=-0.05$	0	-0.0125
$+0.20$	0	0.0500

$\cdot$

$p(x,y)$	$y=0$	0.25
$x=-0.05$	0.1	0.3
$+0.20$	0.5	0.1

$=$

$=$

$x \cdot y$	$y=0$	0.25	
$x=-0.05$	0	-0.00375	
$+0.20$	0	0.00500	
			0.00125

- Calcolo della covarianza

  $\sigma_{XY} = 0.00125 - (0.1)(0.1) = -0.00875.$ negativa!

Significato

- Una covarianza positiva significa che a valori sopra (sotto) la media di X corrispondono valori sopra (sotto) la media di Y
- Una covarianza negativa significa che a valori sopra (sotto) la media di X corrispondono valori sotto (sopra) la media di Y.
- La forza dell'associazione si valuta meglio con il coefficiente di correlazione

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-0.00875}{(0.122)(0.122)} \simeq -0.58$$

Varianza di una combinazione lineare

- Il ricavo di un portafoglio è $T = 1000[\alpha X + (1 - \alpha)Y]$
- T è una combinazione lineare di X e Y

$$T = c_1 X + c_2 Y$$

con coefficienti $c_1 = 1000\alpha$ e $c_2 = 1000(1 - \alpha)$.

Media e varianza di T

$$E(c_1 X + c_2 Y) = c_1 E(X) + c_2 E(Y)$$
$$\sigma^2(c_1 X + c_2 Y) = c_1^2 \sigma_X^2 + c_2^2 \sigma_Y^2 + 2\, c_1 c_2 \sigma_{XY}$$
$$\sigma(c_1 X + c_2 Y) = \sqrt{c_1^2 \sigma_X^2 + c_2^2 \sigma_Y^2 + 2\, c_1 c_2 \sigma_{XY}}$$

Diversificare con $\alpha = 0.2$

- Investiamo $c_1 = 200$ euro al tasso X e $c_2 = 800$ euro al tasso Y: $T = 200X + 800Y$
- $E(T) = (200)(0.1) + 800(0.1) = 100$ euro.
- $\sigma_T^2 = (200^2)(0.015) + (800^2)(0.015) + 2(200)(800)(-0.00875) = 7400$
- $\sigma_T = \sqrt{7400} \simeq 86$ euro.
- Stesso ricavo atteso, ma rischio minore di quello di investire tutto su X o Y: $\sigma(1000X) = 122$ euro

Qual è la diversificazione ottimale?

- Si dimostra che il valore di α che minimizza il rischio è

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

- Risulta nel nostro caso

$$\alpha = \frac{0.015 - (-0.00875)}{0.015 + 0.015 - 2(-0.00875)} = 0.5$$

- Per esercizio mostrate che in tal caso $E(500X + 500Y) = 100$, ma il rischio diventa $\sigma_T = 55.9$ euro.